Woche 12 Anwendungen: Topic Modelling (Teil 1)

Skript

Erarbeitet von
Dr. Jacqueline Klusik-Eckert im Interview mit M. Ed. Stefan Reiners-Selbach

Inhalt

Lernziele

- Wissen über den Einsatz von Clustering als Methoden in den Geisteswissenschaften erhalten
- Unterschiedliche methodische Ansätze für den Einsatz von Clustering kennenlernen
- Kennenlernen eines möglichen Szenarios für den Einsatz von Clustering in den Geisteswissenschaften

Interview

Jacqueline:
Jetzt stehen wir vor einem Haufen mit unglaublich vielen Daten. Textdaten, Bilddaten, Daten über Daten, Metadaten.
Um da etwas herausfinden zu können, muss ich mir das Ganze erstmal ordnen. Ich möchte mir den Haufen sortieren. Am liebsten würde ich das die Mainzelmännchen machen lassen. Das geht aber nicht, denn die gibt es nicht. Deswegen nehme ich das, was es gibt: Künstliche Intelligenz.

Ich habe mir heute einen Spezialisten eingeladen, der uns mal anhand von Beispielen zeigt, wie man Clustering als Methode in den Geisteswissenschaften einsetzen kann. Stefan ist jetzt kein Informatiker, sondern der Koordinator der Digital Humanities an der Heinrich-Heine-Universität Düsseldorf. Hallo Stefan! Digital Humanist, was ist das eigentlich?

Stefan:
Hi Jacqueline, ja danke für die Einladung! Ja, Digital Humanist, das ist eine gute Frage. Also es ist jemand, der in den Digital Humanities arbeitet, der oder die eben digitale Methoden, also zum Beispiel KI-Methoden oder, wo wir heute darüber reden wollen, das Clustering beispielsweise in den Geisteswissenschaften einsetzt.

Jacqueline:
Und wie kann ich mir das jetzt vorstellen, dieses Einsetzen des Clusterings als Methode. Hast du da einfach einen großen Haufen von Daten und denkst dir: „Ja, ok, lass' mal sortieren…" Also, wie benutzt du dieses Verfahren und für was?

Stefan:
Das ist auf jeden Fall auch eine Möglichkeit dafür. Also gerade in den Geisteswissenschaften haben wir es ja viel mit sogenannten retrodigitalisierten Daten zu tun. Also das heißt Daten, die ursprünglich mal nicht als Daten vorgelegen haben, sondern beispielsweise als Bücher im Fall von Textdaten. Das sind dann solche großen Datenmengen, die einfach schon digitalisiert sind und wo wir einen Einblick gewinnen wollen. Worum geht es überhaupt in diesen Daten? Was lassen sich für Themengruppen beispielsweise in diesen Textdaten finden? Und da ist es dann häufig so, dass man einfach vor so einem Berg an Daten steht und Muster darin erkennen möchte, die Daten explorieren, also erkunden möchte, und eben herausfinden möchte, worum geht es darin. Auf der anderen Seite glaube ich, dass man meistens schon eine gewisse Idee hat, womit man sich gerade beschäftigt, weil man die Daten, mit denen man gerade zu tun hat, die hat man ja nicht ohne Grund ausgewählt. Meistens hat man schon eine grobe Vorstellung, worum es gehen sollte, und dann kann man eben hypothesengeleitet vorgehen und Fragen an die Daten stellen und dann eben genauer schauen, ob sich diese Fragen, diese Hypothesen bestätigen oder eben nicht.

Jacqueline:
Gibt es auch bei dir die andere Variante? Also, dass du eine Hypothese hast und sie mittels Clustering bestätigen möchtest?

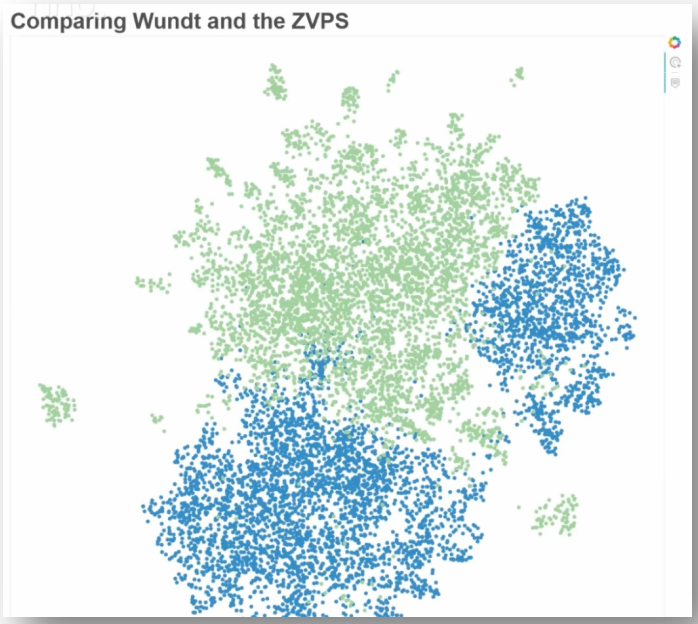
Stefan:
Ja genau, das gibt es eben auch, also das Explorative ist nicht das Einzige. Ich glaube, dass man meistens vom Explorativen auch dazu kommt, Hypothesen zu formulieren und dann vielleicht nochmal genauer nachzuschauen, vielleicht auch mit anderen Methoden als Clustering. So eine ganz klassische Methode für Clustering, die in den Geisteswissenschaften eben auch schon vor den digitalen Methoden etabliert war, ist in der Stilometrie die Autorschaftszuweisung. Wo man eben schaut: Man hat einen Text, man weiß nicht, wer diesen Text geschrieben hat, hat aber die Vermutung, es könnte Autor*in X sein. Vielleicht ist es auch Autor*in Z. Dann nimmt man Texte von Autor*in X und Autor*in Z und packt die zusammen in ein Clustering mit dem Text, wo man eben nicht weiß, wer ihn geschrieben

hat, und dann schaut man, was clustert zusammen. Wenn der Text mit unbekannter Autor*in dann eben nah an Autor*in Z clustert, dann liegt die Vermutung nahe, dass Autor*in Z eben auch diesen Text geschrieben hat. Das Ganze kann man ebenso von Hand machen. Das ist natürlich ziemlich kompliziert. Aber das Ganze kann man eben heute auch mit digitalen Methoden machen und dann vielleicht weniger mit der Frage nach der Autorschaft, sondern beispielsweise, wenn man Korpora miteinander vergleichen möchte.

Jacqueline:
Kannst du mir nochmal auf die Sprünge helfen: Was genau ist ein Korpus?

Stefan:
Ein Korpus ist eine Textmenge, die ich zu einem bestimmten Thema zusammenstelle. In manchen Fällen ergibt sich das aber auch von selbst, weil ich beispielsweise alle Texte einer Zeitschrift oder alle Texte eines Autors, einer Autorin zu einem bestimmten Thema als Korpus zusammenstellen kann. Aber eigentlich ist es immer so ein wenig, ja, auch Urteil der Person, die einen Korpus als Korpus bezeichnet.

Jacqueline:
Stefan, das ist ja unglaublich spannend, aber grau ist jede Theorie: Kannst du uns denn auch etwas zeigen?



Stefan:
Ja sehr gerne, sogar etwas Buntes. Beim Clustering wird es dann ja gerne bunt. Ich habe euch hier ein erstes Beispiel mitgebracht, nämlich einen solchen Vergleich von zwei Korpora. Also diese Visualisierung ist so zu lesen, dass jeder Punkt einem Text entspricht. Das Verfahren, was diese Punkte herstellt, das ist ein Vektorizer, der also die Texte vektorisiert, in Vektoren umwandelt. Das sind dann hochdimensionale Vektoren. Dann

benutzt man eine Dimensionsreduktion, um das Ganze auf der Ebene in dem zweidimensionalen Raum hier visualisieren zu können. Und das Ganze ist so zu lesen, dass je näher sich zwei Punkte liegen, desto ähnlicher sind die Texte laut Algorithmus. Die Einfärbung, die ihr hier seht, das ist gar kein dezidierter Clustering-Algorithmus gewesen, sondern was ich hier eingefärbt habe, ist quasi die Herkunft der Texte, aus welchem Korpus sie stammen. Jetzt ist es interessant, eben diese Einfärbung mit der Lage der Punkte zu vergleichen. Also wir sehen in der Mitte, dass da ziemlich viele Punkte aus beiden Korpora auf einem Fleck liegen, und das heißt, hier haben wir es mit einer Textmenge zu tun, wo die Themen, die behandelt werden, wohl sehr ähnlich sind, wohingegen es dann solche Randbereiche gibt, wo es dann nahezu gar keine Überschneidungen gibt. Das heißt, die beiden Textkorpora sollen eigentlich dasselbe Thema behandeln, man sieht hier aber ziemlich deutlich, dass es doch ganze Themenkomplexe gibt, die eben überhaupt nicht von beiden behandelt werden.

Jacqueline:
Und auch hier so ein paar kleine Ausreißer, oder?

Stefan:
Und auch einige kleine Ausreißer würde ich sagen. Man sieht hier unten einen Ausreißer und auf der anderen Seite einen Ausreißer, die also völlig unterschiedlich sind, die sich vom Thema also völlig von den anderen Korpora unterscheiden. Hier wird es dann total spannend einzusteigen, was sind das für Texte, warum sind die überhaupt in diesen Korpus gekommen und quasi, was haben die mit den anderen zu tun und an welchen Stellen eben nicht.

Quellen

Quelle [1] Zeitschrift für Völkerpsychologie und Sprachwissenschaft. (1860.). https://opacplus.bsb-muenchen.de/search?id

Weitere verwendete Literatur

Angelov, Dimo (2020): Top2Vec: Distributed Representations of Topics, arXiv:2008.09470v1

Bokeh Development Team (2018). Bokeh: Python library for interactive visualization, http://www.bokeh.pydata.org

McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018

Noichl, Maximilian (2019): Modeling the structure of recent philosophy, Synthese, https://doi.org/10.1007/s11229-019-02390-8

Pence, C. H. (2022). Testing and discovery: Responding to challenges to digital philosophy of science. Metaphilosophy, 53, 238–253. https://doi.org/10.1111/meta.12549

Lean, O. M., Rivelli, L., & Pence, C. H. (2023). Digital Literature Analysis for Empirical Philosophy of Science'. British Journal for the Philosophy of Science, 74, https://doi.org/10.1086/715049

Disclaimer

Transkript zu dem Video „Woche 12 Anwendungen: Anwendung von Clustering in den Geisteswissenschaften (Teil 1)", Dr. Jacqueline Klusik-Eckert, M. Ed. Stefan Reiners-Selbach.